File S1

1. Estimation of parameters

While the Geometric-Poisson distribution appears to approximate the distance distribution under simulation well, this is under the assumption that several key parameters of interest are known – namely, the mutation rate, the equilibrium effective population size within-host, and the bottleneck size. With a known transmission structure (for instance, within a household (COWLING *et al.* 2010)), it is possible to estimate some of these quantities. We simulated an outbreak and assumed that a set of 25 transmission pairs was observed. Figure S8 shows the likelihood of these data under a range of values for mutation rate and effective population size. The estimate of the effective population size is uncertain, since the data are less informative of this parameter; in the most extreme case, where coalescence occurs immediately prior to the time of lineage divergence, the likelihood function depends only on the mutation rate.

The bottleneck size can additionally be estimated. Observation of multiple genotypes shortly after a bottleneck event suggests that the bottleneck must be large enough to allow diversity through; Figure S9 shows the likelihood of observing different numbers of SNPs within host shortly after transmission, for a range of potential bottleneck sizes. Again, such estimates are associated with very high levels of uncertainty, particularly for large bottleneck sizes. However, it may be possible to test the hypothesis that the bottleneck size is strict, an assumption frequently made in transmission network reconstruction methods.

2. Simulated outbreak

Figure S2 shows a simulated SIR outbreak with 25 infected individuals, 18 of which have a sampled genotype. We considered the relative likelihood of observing a genetic distance between two hosts, given direct transmission has occurred (Figure S2, bottom left). The maximum likelihood estimate of transmission source was correct in eight out of 17 transmission events. In comparison, selecting the genetically closest isolate as the source was correct in seven cases, although for some of these, multiple hosts were equally close.

For any given infected host, a genetic distance threshold may be specified, which may be used to rule out direct transmission to a given probability level. Consider the individual labelled 'N' in figure S2, with a sample at time 1000. Under the geometric-Poisson approximation with strict bottleneck, the probability of drawing a sample differing by 4 SNPs or greater at time 1000 from the true host is less than 5%. As such, six of the eleven previously infected individuals can be ruled out as transmission sources at this level. As the time between samples and/or the bottleneck size increase, this threshold also increases.

3. Comparison with transmission network estimation software packages.

'*Outbreaker*' is an R package for the investigation of individual-level transmission dynamics using genomic data (JOMBART *et al.* 2014), while '*seqTrack*' is an earlier and simpler method, implemented in the '*adegenet*' package (JOMBART *et al.* 2011). These software packages are arguably the most accessible tools for estimating a transmission network available at present, and as such, we wanted to compare their performance against our method. Given a user-specified infectivity distribution and one genomic sample per infected host, *outbreaker* implements an MCMC algorithm which estimates the posterior edge probabilities of the network, along with several parameters of interest, including the mutation rate. Unlike our model, this approach therefore does not require infection times and mutation rate to be known (and can also be used to detect importations into a population), however, it operates on a less sophisticated model of within-host dynamics – mutations are assumed to be a feature of transmission, and an infected host is adequately represented by a single sequenced pathogen isolate. *seqTrack* identifies the genetically closest pathogen sample as the source, using the specified mutation rate to break ties. This approach also assumes that each host is represented by one genomic sample.

We simulated outbreaks under various assumptions, and attempted to identify the transmission network using our likelihood approach, as well as the *outbreaker* and *seqTrack* functions. While the *outbreaker* package can also be used to simulate outbreaks, this is performed under the assumptions mentioned previously, so we instead simulated the within-host pathogen dynamics explicitly, as described in *Methods*. We used the number of transmission routes to compare the two methods. We ran *outbreaker* with no spatial model, and detection of importations suppressed. Furthermore, we assumed a flat infectivity distribution. We emphasize that these approaches are not directly comparable, since *outbreaker* and *seqTrack* accommodate unknown infection times, and *outbreaker* furthermore estimates the mutation rate, giving our approach an advantage in this comparison. Results are presented in Table S2.

4. MRSA outbreak analysis

While the analysis provided in the main text provides estimates of transmission routes under plausible parameter values found in the literature, there is a great deal of uncertainty surrounding true within-host pathogen population dynamics, and as such, we repeated the analysis under a range of assumptions. The mutation rate used in the main analysis was given in the paper describing this dataset; the mutation rate of MRSA has previously been estimated to be higher (3×10^{-6} per nucleotide per year, equivalent to 5×10^{-4} per genome per generation (Harris *et al.* 2010; Young *et al.* 2012)), so we repeated the analysis with this value. With this higher mutation rate, a larger range of genetic distances are plausible, and as such, fewer routes were excluded at the 5% level. The HCW was a plausible source for most patients on the ward, however, the genetic distance from patients 1 and 5 to the HCW were more similar than would be expected, given this infection route. No patient to HCW transmission route could be excluded at the 5% level.

Changing the effective population size had a limited effect on the estimated transmission route estimates. Values of 2000 and higher produced near identical posterior probabilities. Previous studies have estimated nasal carriage of *S. aureus* to have an effective population size in the range of 50-4000 (Young *et al.* 2012; Golubchik *et al.* 2013). We experimented with an effective population size of 100, finding that five patient-HCW routes, and seven HCW-patient routes could be excluded at the 5% level.

Varying the time at which the HCW became infected had an impact on posterior transmission probabilities. Moving this value forward in time decreases the number of SNPs expected to accumulate by the time of observation. If the HCW infection time was 164 days after the first case, the upper bound of the range provided by (Harris *et al.* 2013), five patients remain temporally consistent with having become infected by the HCW. Two of these transmission routes can be excluded at the 5% level.

We repeated our analysis using the pure Poisson model. In general, this distribution has a shorter right tail than the geometric-Poisson distribution, and as such, can lead to more transmission routes being rejected at a given probability level. With the same assumptions as in the main text, the HCW-patient routes were typically given a higher posterior probability under the Poisson distribution, however, the most likely source of infection remained the same for all individuals (Figure S5).

5. Conditional distributions

We define a phylogenetic subtree to be the unique set of branch segments linking two isolates, originating at the time of their coalescence. Then the genetic distance $\psi(g_1,g_2)$ is dependent on another distance $\psi(g_3,g_4)$ by the intersection of the two phylogenetic subtrees. The conditional distribution of one genetic distance given another is

$$\begin{aligned}\psi(g_1,g_2)\,|\,\psi(g_3,g_4) \sim \mathrm{Bin}\left(\psi(g_3,g_4),\frac{\text{length of intersection}}{\text{length of subtree}(g_3,g_4)}\right)\\ +\,\mathrm{Pois}\{\mu((\text{length of subtree}(g_1,g_2))-(\text{length of intersection}))\}\end{aligned} \tag{8}$$

Figure S7 shows two possible configurations of the phylogenetic and transmission tree with three infected cases. In both settings, $\psi(g_2,g_3)$ depends on $\psi(g_1,g_2)$ via the mutations occurring along branch b_3. If the sequences at the internal nodes are known, or can be inferred, this estimation is unnecessary, as the true number of mutations along any given branch segment can be calculated. However, since the genealogy is not typically observed, and does not necessarily correspond to the transmission network, even under a strict bottleneck (PYBUS and RAMBAUT 2009; YPMA *et al.* 2013), such an approximation may be useful for inference of the full network, and to account for multiple samples per host.

Transmission chains of length 3 were simulated to investigate conditional distributions of genetic distances. Times from infection to sampling and onward transmission were identical for all cases. With a strict bottleneck, $\psi(g_2,g_3)$ varies only minimally with $\psi(g_1,g_2)$, but $\psi(g_1,g_3)$ shows a clear dependency. Both distances increase with greater values of $\psi(g_1,g_2)$ under larger bottlenecks (Figure S6). With a strict bottleneck, the scenario in Figure S7B is impossible, and as such, the intersection of subtrees (g_1,g_2) and (g_2,g_3) is relatively small. With an increasing bottleneck size, the probability of scenario B, and therefore the potential length of subtree overlap, increases.

References

COWLING, B. J., K. H. CHAN, V. J. FANG, L. L. H. LAU, H. C. SO *et al.*, 2010 Comparative Epidemiology of Pandemic and Seasonal Influenza A in Households. New England Journal of Medicine 362: 2175-2184.

GOLUBCHIK, T., E. M. BATTY, R. R. MILLER, H. FARR, B. C. YOUNG *et al.*, 2013 Within-Host Evolution of Staphylococcus aureus during Asymptomatic Carriage. PLoS One 8: e61319.

HARRIS, S. R., E. J. P. CARTWRIGHT, M. E. TÖRÖK, M. T. G. HOLDEN, N. M. BROWN *et al.*, 2013 Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infectious Diseases 13: 130-136.

HARRIS, S. R., E. J. FEIL, M. T. G. HOLDEN, M. A. QUAIL, E. K. NICKERSON *et al.*, 2010 Evolution of MRSA during hospital transmission and intercontinental spread. Science 327: 469-474.

JOMBART, T., A. CORI, X. DIDELOT, S. CAUCHEMEZ, C. FRASER *et al.*, 2014 Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLoS Computational Biology 10: e1003457.

JOMBART, T., R. M. EGGO, P. J. DODD and F. BALLOUX, 2011 Reconstructing disease outbreaks from genetic data: a graph approach. Heredity 106: 383-390.

PYBUS, O. G., and A. RAMBAUT, 2009 Evolutionary analysis of the dynamics of viral infectious disease. Nature Reviews Genetics 10: 540-550.

YOUNG, B. C., T. GOLUBCHIK, E. M. BATTY, R. FUNG, H. LARNER-SVENSSON *et al.*, 2012 Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. PNAS 109: 4550-4555.

YPMA, R. J. F., W. M. VAN BALLEGOOIJEN and J. WALLINGA, 2013 Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 195: 1055-1062.